第7章

Management of Experimental Data in Analytical Chemistry

分析化学中的数据处理

7.1 统计学(statistics)中常用术语

1、总体：

所考察的对象的全体。

2、样本（子样）：

自总体中随机抽出的一组测量值。

3、样本大小（或容量）：

样本中所含测量值的数目。

7.2 分析化学中常用统计量

1、总体平均值：

样本平均值为
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

当 n 无限增大时，得总体平均值 μ

$$\mu = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i$$

若无系统误差

$$\mu = x_T$$

2、总体平均偏差：

样本平均偏差为 $$\bar{d} = \frac{\sum |x - \bar{x}|}{n}$$

当 n 无限增大时，得总体平均偏差 δ

$$\delta = \frac{\sum |x - \mu|}{n}$$

3、 样本(sample)标准偏差：

衡量测量值的分散程度，即精密度的好坏。

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

式中(n - 1) 称为自由度，以 f 表示。

自由度表明独立偏差的个数。

当 n 很大时，$\bar{x} \to \mu \quad s \to \sigma$

4、总体标准偏差 σ

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

5、总体标准偏差与总体平均偏差

当 n >20 时

$$\delta = 0.7979\sigma \approx 0.80\sigma$$

6、平均值的标准偏差：

总体：

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

样本（有限次测量）：

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

平均值的标准偏差与测定次数的平方根成反比

统计量符号总结

	平均值	单次测量的平均偏差	标准偏差	平均值的标准偏差
样本	$\overline{x}$	$\overline{d}$	s	$s_{\overline{x}}$
总体	μ	δ	σ	$\sigma_{\overline{x}}$

7.3 少量数据的统计处理

Statistics for Small Sets of Numbers

7.3.1 **t** 分布曲线

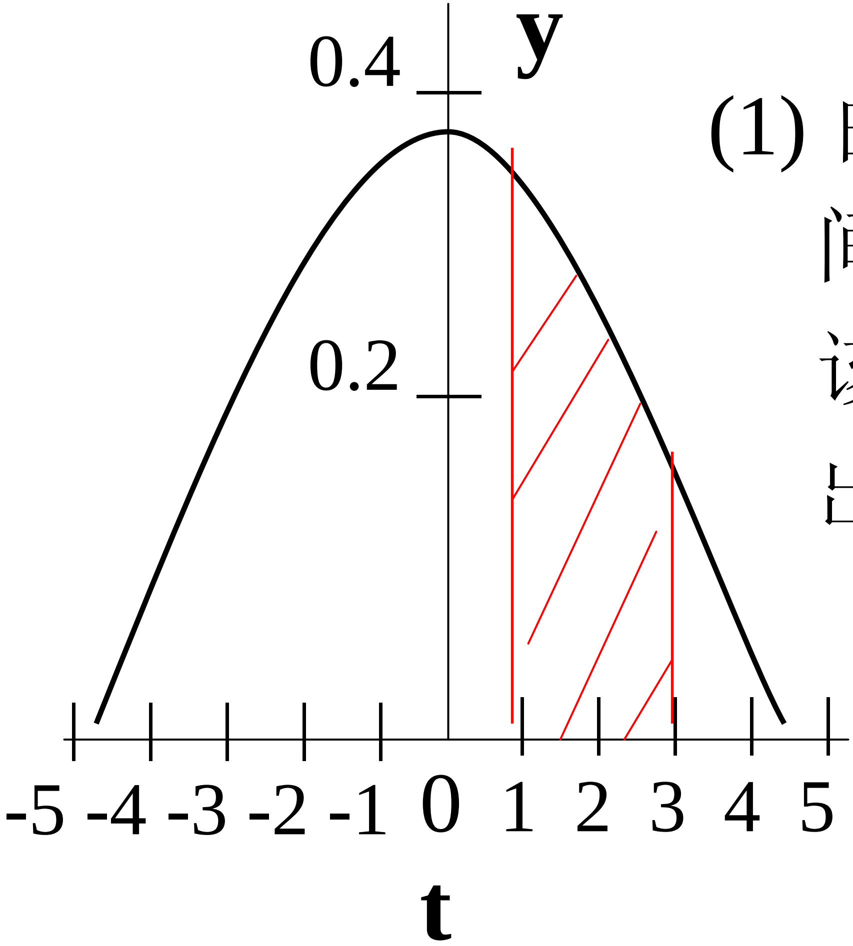
1、**t** 的定义

$$t = \frac{x-\mu}{s} \quad 或 \quad t = \frac{\bar{x}-\mu}{s_{\bar{x}}}$$

↓

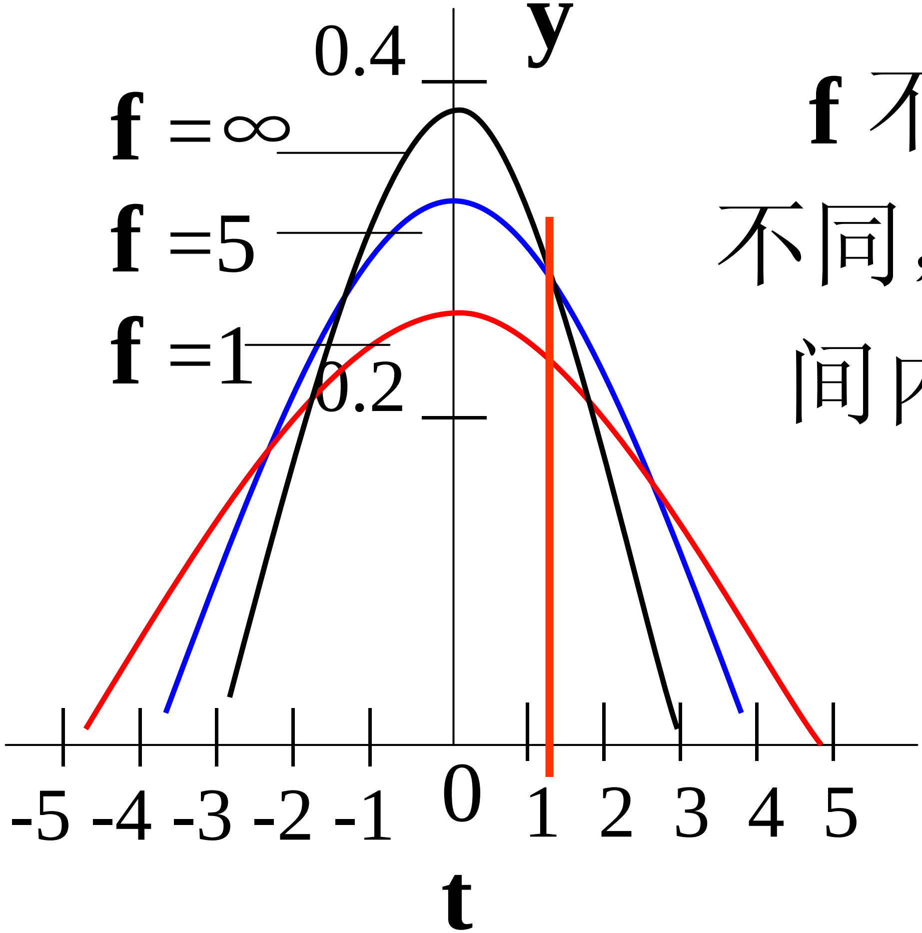
$x = \mu \pm t \cdot s$ 表示测定值出现的范围。

2、**t** 分布曲线的形状与特点



(1) 曲线下一定 **t** 值区间内的积分面积是该区 间内随机误差出现的概率 **P**。

(2) 曲线形状与 **t** 值和自由度 **f** 有关。



f 不同，曲线形状不同，**t** 值一定的区间内随机误差出现的概率 **P** 不同。

3、 $t_{\alpha,f}$ 值表：体现 **t , f , P** 之间的关系

表中置信度就是概率 **P** ，

它表示 **t** 值一定时，测定值落在 $(\mu \pm t \cdot s)$ 范围内的概率。

显著性水准 $\alpha = 1 - P$

自由度 $f = n - 1$， n 为测定次数

双边：表示**P**是从**-t**到**+t**积分而得的面积。

已知 **f , P** 时，可从表中查出相应的 **t** 值。

7.3.2 平均值的置信区间

1、置信区间的计算公式

根据

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}}$$

可得计算公式

$$\mu = \overline{x} \pm t \cdot s_{\overline{x}} = \overline{x} \pm \frac{ts}{\sqrt{n}}$$

2、置信区间的概念

$$\mu = \bar{x} \pm t \cdot s_{\bar{x}} = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

它表示在一定置信度下，以平均值 $\bar{x}$ 为中心，包括总体平均值 μ 的范围。

μ 是客观存在的，没有随机性，不能说它落在某一区间的概率是多少；只能说某区间包括总体平均值的概率是多少。

3、平均值置信区间的计算步骤

根据
$$\mu = \bar{x} \pm \frac{t_{\alpha,f} \cdot s}{\sqrt{n}} \qquad （1）$$

①计算测量值的 $\bar{x}$，s

②由置信度和测量值个数 n，计算出

$\alpha = 1 - P$ ，$f = n - 1$ 从表 7-3 查 $t_{\alpha,f}$

③将上述数据代入(1)式即可计算。

分析化学中一般将置信度定在**95%**或**90%** 。

平均值置信区间计算示例：P268习题No.7

解：（1）计算平均值和标准偏差

$$\bar{x} = (0.160 + 0.152 + 0.155 + 0.154 + 0.153 + 0.156) \div 6 = 0.155 \ mg \cdot mL^{-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{6}(x_i - 0.155)^2}{6-1}} = 0.0028$$

- 因为 P = 95%， n = 6

所以 $\alpha = 1 - P = 1 - 0.95 = 0.05$

$$f = n - 1 = 6 - 1 = 5$$

查表 7-3 得 $t_{0.05,5} = 2.57$

计算得

$$\mu = 0.155 \pm \frac{2.57 \times 0.0028}{\sqrt{6}}$$

$$= 0.155 \pm 0.003\, mg \cdot mL^{-1}$$

7.3.3 显著性检验

1、概念及作用

测定标准试样

$\bar{x} \sim \mu$ 有差异

同一试样的两组分析结果

$\bar{x}_1 \sim \bar{x}_2$ 有差异

假设检验

存在显著性差异 有系统误差

不存在显著性差异 无系统误差

2、显著性检验的方法：

（1） **t** 检验法： 检验 $\bar{x} \sim \mu$

检验平均值与标准值之间的差异。

方法原理　　假设 $\bar{x}$ 与 μ 无显著性差异，则符合关系式

$$\mu = \bar{x} \pm \frac{t \cdot s}{\sqrt{n}} \qquad (1)$$

由(1)式计算出的 $\mathbf{t}_{计}$ 应等于或小于 $\mathbf{t}_{表}$。

方法原理(接上页)

若 $\mathbf{t}_{计} \leqslant \mathbf{t}_{表}$，与假设相符，无显著性差异

若 $\mathbf{t}_{计} > \mathbf{t}_{表}$，与假设不符，存在显著性差异

t 值计算公式 由（1）式可得

$$t = \frac{\left| \bar{x} - \mu \right|}{s} \cdot \sqrt{n}$$

检验时的标准置信度

$$P = 95\%$$

检验步骤

a. 根据题意计算 $\mathbf{t}_{计}$

b. 计算出 α , f , 查 $\mathbf{t}_{表}$

c. 比较 $\mathbf{t}_{计}$与 $\mathbf{t}_{表}$， 然后作出结论

t 检验法示例　P269　No. 12

解：已知 $\mu = 54.46\%$ ， $\bar{x} = 54.26\%$

$$s = 0.05\%, n = 4, P = 95\%$$

计算 **t**

$$t = \frac{|\bar{x} - \mu|}{s} \cdot \sqrt{n} = \frac{|54.26\% - 54.46\%|}{0.05\%} \cdot \sqrt{4}$$

$$\mathbf{t}_{计} = 8$$

计算 α , f ,

$$\alpha = 1 - P = 1 - 0.95 = 0.05$$

$$f = n - 1 = 4 - 1 = 3$$

查表7-3，得 $t_{0.05,3} = 3.18$

因为 $\mathbf{t}_{计} = 8 > \mathbf{t}_{0.05,3}$

所以，平均值与标准值存在显著性差异。

分析结果存在系统误差。

（2）**F 检验法：**

通过比较两组数据的方差 s^2，确定其精密度(标准偏差 s_1 与 s_2)是否有显著性差异的方法。

统计量 F 的定义与计算：

$$F = \frac{s_{大}^2}{s_{小}^2}$$

大：大方差

小：小方差

- 由**F** 的定义式可计算得**F** 值

若**F**→1，则 S_1 与 S_2 相差不大；

若 **F** 较大，则 S_1 与 S_2 存在显著性差异。

检验步骤

a. 计算 **F** 值，

b. 计算 $f_{大}$，$f_{小}$，查**F** 值表得$F_{表}$

c. 比较：若 $F_{计}>F_{表}$，存在显著性差异；

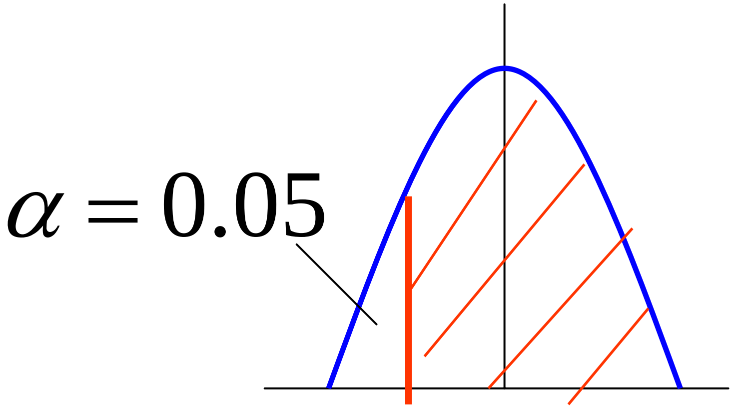
反之，不存在显著性差异。

F 检验的单、双边问题

若被检验的两组数据的精密度可能是

$s_1 > s_2$ 或 $s_1 = s_2$，

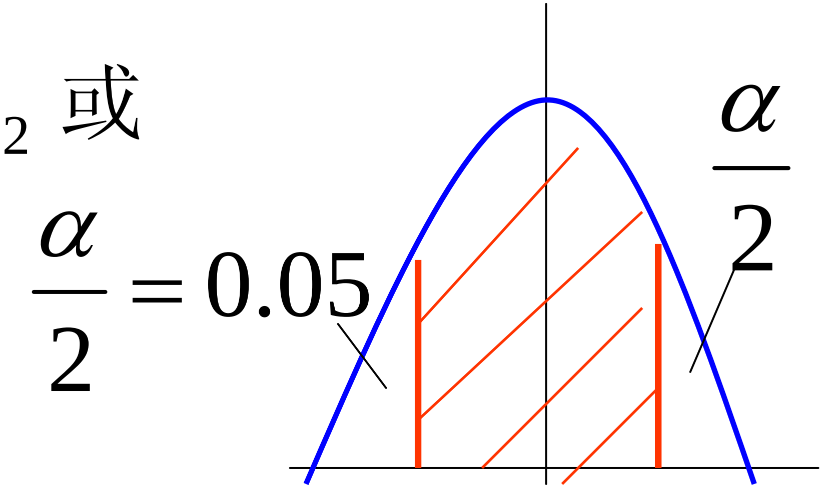
则为单边问题， P=95%



若精密度可能是 $s_1 > s_2$ 或 $s_1 = s_2$ 或 $s_1 < s_2$，

则为双边问题，

P=90%



F 检验法示例　P269 No. 11

解：已知置信度为90%，因此是双边问题。

(1) 计算得

$$\bar{x}_A = 9.56\ ,\ s_A = 0.057$$

$$\bar{x}_B = 9.47\ ,\ s_B = 0.085$$

(2)计算 **F**

$$F = \frac{s_{大}^2}{s_{小}^2} = \frac{(0.085)^2}{(0.057)^2} = 2.22$$

(3) 查 **F** 值表：

计算自由度

$$f_{大} = f_{小} = 6 - 1 = 5$$

由表7-4 查得 $\mathbf{F}_{表} = 5.05$

(4) 由于 $\mathbf{F}_{计} = 2.22$， 故 $\mathbf{F}_{计} < \mathbf{F}_{表}$

所以，该两组数据的精密度无显著性差异。

作出此判断的置信度为**90%** 。

（3） **F** 检验法 + **t** 检验法

设两组分析数据的测定次数、标准偏差和平均值分别为 $n_1 \quad s_1 \quad \bar{x}_1$

$$n_2 \quad s_2 \quad \bar{x}_2$$

作用： **F 检验法 + t 检验法**可检验两组分析结果的**平均值** $\bar{x}_1, \bar{x}_2$ 之间是否有显著性差异。

检验步骤：

①用**F** 检验法： 检验 S_1与 S_2之间是否存在显著性差异。

如 S_1与 S_2之间存在显著性差异，则两组分析数据存在显著性差异。

如 S_1与 S_2之间不存在显著性差异，则可认为$S_1 \approx S_2$，可计算合并标准偏差继续进行下述检验。

②计算合并标准偏差

因为

$$s=\sqrt{\frac{\text{偏差平方和}}{\text{总自由度}}}$$

所以计算公式为

$$s=\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1-1)+(n_2-1)}}$$

③用 **t** 检验法

计算 **t** 值

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

计算总自由度 $f = n_1 + n_2 - 2$

计算显著性水准 $\alpha = 1 - P$

查 **t** 值表，得 $\mathbf{t}_{表}$ 值。

若 $\mathbf{t}_{计} > \mathbf{t}_{表}$

则两组数据的平均值存在显著性差异。

若 $\mathbf{t}_{计} < \mathbf{t}_{表}$

则两组数据的平均值不存在显著性差异。

请见显著性检验示例 例5，6，7，8，9

F 检验法 + t 检验法 示例 **P269 No.15**

解：$\bar{x}_A = 60.45\%, s_A = 0.048\%, n_A = 4$

$\bar{x}_B = 60.11\%, s_B = 0.051\%, n_B = 4$

作F 检验

$$F = \frac{s_{大}^2}{s_{小}^2} = \frac{(0.051\%)^2}{(0.048\%)^2} = 1.13$$

由 $f_{大} = f_{小} = 4-1 = 3$ 查得 $\mathbf{F}_{表} = \mathbf{9.28}$

因为 $\mathbf{F}_{计} < \mathbf{F}_{表}$，故精密度间无显著性差异。

- 计算合并标准偏差

$$s=\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1-1)+(n_2-1)}}$$

$$=\sqrt{\frac{(0.048\%)^2(4-1)+(0.051)^2(4-1)}{(4-1)+(4-1)}}$$

$$s=0.050\%$$

- 作 **t** 检验

计算
$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

$$= \frac{\left|60.45\% - 60.11\%\right|}{0.050\%}\sqrt{\frac{4 \times 4}{4 + 4}}$$

$$t = 9.62$$

- 总自由度 $f = 4 + 4 - 2 = 6$

置信度 $P = 90\%$ $\alpha = 1 - 0.90 = 0.10$

查 **t** 值表得 $t_{0.10,6} = 1.94$

因为 $\mathbf{t}_{计} = \mathbf{9.62}$ 得 $\mathbf{t}_{计} > \mathbf{t}_{表}$

所以两瓶试剂含 **Cl⁻** 的质量分数

有显著性差异。

7.3.4 异常值的取舍

异常值　一组数据中离群较远的数据，称为异常值或可疑值或极端值。

异常值的取舍

过失异常——必须舍去。

非过失异常——用统计学方法处理。

统计学处理异常值的方法

$4\overline{d}$ 法，Q检验法，格鲁布斯法

1、$4\bar{d}$ 法

根据正态分布规律

$$|x-\mu| > 3\sigma \qquad P < 0.3\%$$

因为 $\delta = 0.80\sigma$ 所以 $4\delta \approx 3\sigma$

对于少量数据，只能用 $\bar{d}$ 代替 δ

只要 $|x-\bar{x}| > 4\bar{d}$ 则 x 可舍去。

可见本方法较简单，但误差较大。

- $4\bar{d}$ 法判断异常值取舍的步骤
- 例 P270, No.17第一问

解： ①求出除 0.1016 之外的其余数据的平均值和平均偏差：(单位略去)

$$\bar{x} = (0.1011 + 0.1010 + 0.1012) \div 3 = 0.1011$$

$$\bar{d} = \frac{|0.1011 - \bar{x}| + |0.1010 - \bar{x}| + |0.1012 - \bar{x}|}{3}$$

$$= 0.00007$$

- ②求出 $4\bar{d}$

  因为 $\bar{d} = 0.00007$

  所以 $4\bar{d} = 4 \times 0.00007 \approx 0.0003$

③计算被判断值 0.1016 与平均值的偏差：

$$\left| x - \bar{x} \right| = \left| 0.1016 - 0.1011 \right| = 0.0005$$

④判断：$\left| x - \bar{x} \right| = 0.0005 > 4\bar{d} = 0.0003$

所以 第 4 次数据 0.1016mol.L^{-1} 应舍去。

3、格鲁布斯(Grubbs)法

方法步骤

① 将一组数据从小到大排列

$$x_1, x_2, \ldots, x_{n-1}, x_n$$

x_1和 x_n 是极端值，可能是异常值。

② 计算全组数据的平均值和标准偏差

即 $\bar{x}$，s

③计算统计量 T：

若判断 $\mathbf{x_1}$，则 $$T = \frac{\bar{x} - x_1}{s}$$

若判断 $\mathbf{x_n}$，则 $$T = \frac{x_n - \bar{x}}{s}$$

④查 $T_{\alpha,n}$ 值表 (P 256, 表 7-5)

已知显著性水准和测定次数，就可得 **T**.

⑤比较：如 $T_{计} > T_{\alpha,n}$ 则异常值应舍去。

否则应保留。

格鲁布斯法示例（P270 No.18 第一问）

解：从小到大排列(单位略去)

0.1010， 0.1011 , 0.1012 . 0.1016

计算平均值和标准偏差

$$\bar{x} = (0.1010 + 0.1011 + 0.1012 + 0.1016) \div 4$$
$$= 0.1012$$

$$s = \sqrt{\frac{\sum_{i=1}^{4}(x_i - 0.1012)^2}{4-1}} = 0.00026$$

- **计算T值**： 0.1010和 0.1016是两个极端值。

$$T_n = \frac{x_n - \bar{x}}{s} = \frac{0.1016-0.1012}{0.00026} = 1.54$$

$$T_1 = \frac{\bar{x} - x_1}{s} = \frac{0.1012-0.1010}{0.00026} = 0.77$$

置信度为 95% , $\alpha = 1-0.95 = 0.05$， $n = 4$

查表 7-5 $T_{0.05,4} = 1.46$

- 比较：$$T_n = 1.54 > T_{0.05,4} = 1.46$$

所以 0.1016 应舍去。

$$T_1 = 0.77 < T_{0.05,4} = 1.46$$

所以 0.1010 应保留。

格鲁布斯法准确性较好，如与其它方法有矛盾时，应以格鲁布斯法为准。

3、Q 检验法

方法步骤

①将一组数据从小到大排列

$$x_1, x_2, \ldots, x_{n-1}, x_n$$

x_1和 x_n 是极端值，可能是异常值。

②计算统计量Q

如判断$\mathbf{x_n}$，则
$$Q = \frac{x_n - x_{n-1}}{x_n - x_1}$$

- 如判断 $\mathbf{x_1}$，则

$$Q = \frac{x_2 - x_1}{x_n - x_1}$$

Q值越大，说明 x_1或 x_n离群越远。

Q也称舍弃商。

③查Q值表：

由置信度P和测定次数n可查Q值。

④比较：如 $Q_{计} > Q_{表}$，则异常值应舍去，

反之，应保留。

- Q检验法示例（用P270 No.17中的数据）

解：从小到大排列(单位略去)

0.1010，0.1011 , 0.1012 . 0.1016

计算**Q**值：分别判断 0.1010 和 0.1016

$$Q_1 = \frac{x_2 - x_1}{x_n - x_1} = \frac{0.1011 - 0.1010}{0.1016 - 0.1010} = 0.17$$

$$Q_n = \frac{x_n - x_{n-1}}{x_n - x_1} = \frac{0.1016 - 0.1012}{0.1016 - 0.1010} = 0.67$$

查Q值表7-6：

已知 n = 4， 设置信度 P=90%

查得 $$Q_{0.90,4} = 0.76$$

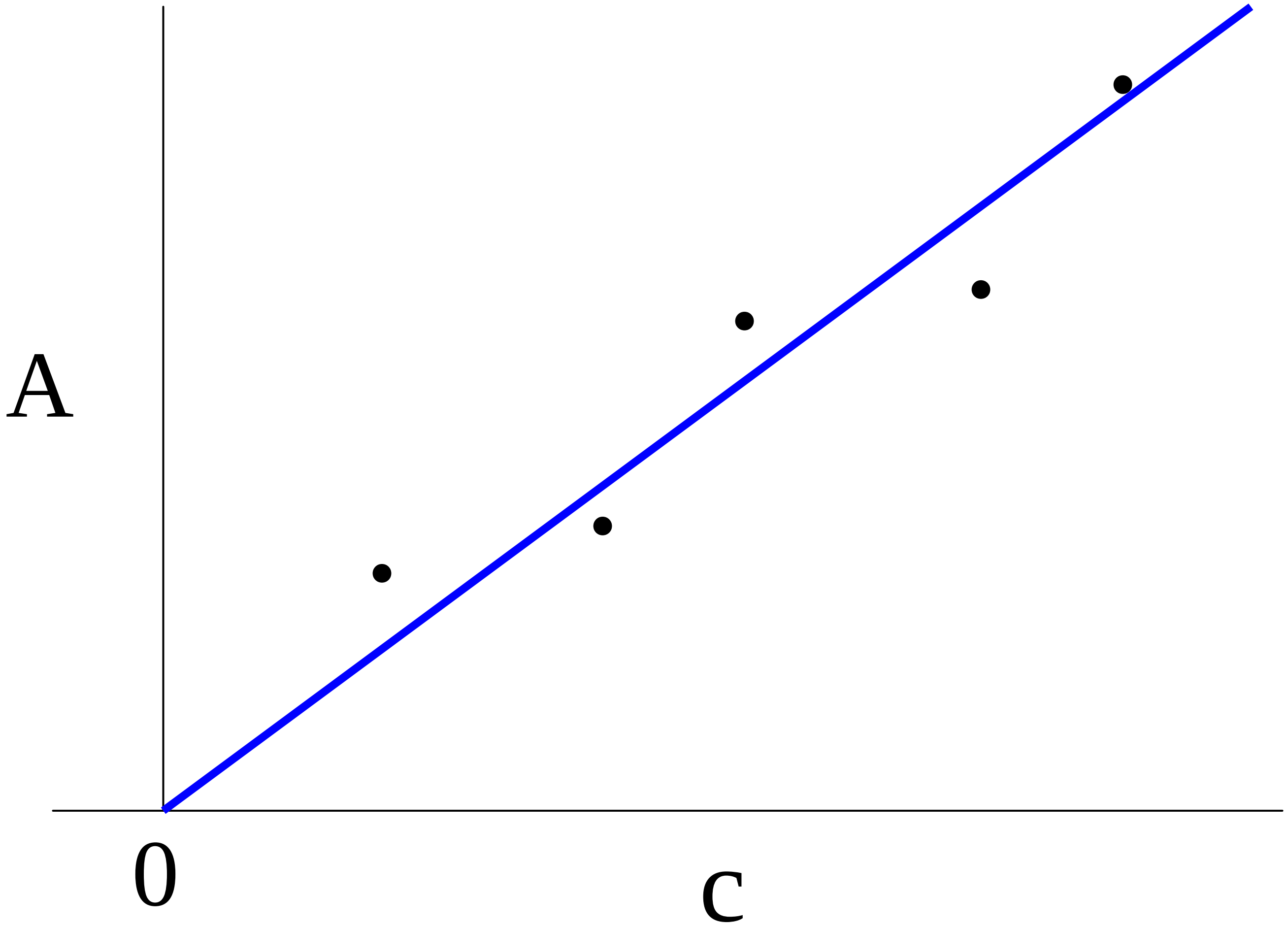
比较 $$Q_1 = 0.17 < Q_{0.90,4} = 0.76$$

$$Q_n = 0.67 < Q_{0.90,4} = 0.76$$

所以，0.1010 和 0.1016 都应保留。

7.4 回归分析法(Regressive Analysis)

用标准曲线来获得未知物质含量时，通过数理统计方法找出对各数据点误差最小的直线，并且估计直线上各点的精密度以及数据间的相关关系的方法，属于回归分析的问题。


A
0
c

7.4.1 一元线性回归方程

1、基本概念

最小二乘法：数理统计中使直线上所有测量值(y)的残(偏)差平方和为最小的方法。

回归直线：

使用最小二乘法通过测量点所确立的最能反映其真实分布状况的最佳直线。

一元线性回归方程：

分析化学中的标准曲线(校正曲线)可用一元线性方程表示，该标准曲线用最小二乘法处理后的回归直线方程

$y = a + bx$ 称为一元线性回归方程。

回归系数：

一元线性回归方程 $y = a + bx$ 中的系数 a、b 称为回归系数。

2、回归系数 a、b 的确定

设通过 n 个实验点(x_i,y_i)的校正曲线为

$$y_i = a + bx_i + e_i$$ 式中e_i为残(偏)差

设校正曲线不存在残差的方程为

$$\hat{y} = a + bx_i$$

令残差平方和为Q

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

- 如能**使残差平方和Q达最小**，就能得到一条对各数据点误差最小的校正曲线。
- 故用 Q 对a、b 求偏微商并令其等于零

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$$

- 解上述两偏微分方程得到回归系数：

a 的计算公式

$$a=\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}=\bar{y}-b\bar{x}$$

- **回归系数 b 的计算公式**

$$b = \frac{\sum_{i=1}^{n} [(x_i - \overline{x})(y_i - \overline{y})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

根据测量值计算出 a, b，即可确定一元线性回归方程为　　y = a + bx

- 7.4.2 相关系数 r

用以检验回归直线上的两个变量(x , y)之间的线性关系是否有意义的系数。

1、相关系数的物理意义

(1) $r = 1$：所有的 y_i 值都在回归线上。

(2) $r = 0$： y 与 x 之间完全不存在线性关系。

(3) $0 < r < 1$： y 与 x 之间存在线性关系，r 值越接近1，线性关系就越好。

2、相关系数 r 的定义式

$$r = b\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

式中 b 为回归系数

相关系数 r 的定义式

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

3、检验方法

（1）根据 r 的定义式计算 $r_{计}$，

（2）根据置信度及自由度查表7-7得 $r_{表}$，

（3）比较：

若 $r_{计} > r_{表}$， y 与 x 的线性关系有意义

反之， y 与 x 的线性关系没有意义

r 愈接近 1， y 与 x 的线性关系愈好。

- 回归分析法示例：P271 No.29

解：（1）$\overline{x} = 0.60\,\mathrm{m}g, \qquad \overline{y} = 0.178$

$$\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y}) = 0.102$$

$$\sum_{i=1}^{5}(x_i - \overline{x})^2 = 0.40$$

- 回归系数 b

$$b=\frac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{5}(x_i-\overline{x})^2}=\frac{0.102}{0.40}=0.255$$

- 回归系数 a

$$a = \overline{y} - b\overline{x} = 0.178 - 0.255 \times 0.40 = 0.025$$

一元线性回归方程为

$$y = 0.025 + 0.255\,x$$

(2)未知液中含**Fe**量 x

将未知液的吸光度 **y=0.205**代入回归方程

得 $$0.205 = 0.025 + 0.255\,x$$

$$x = 0.71mg$$

(3)计算相关系数 r

$$r = b\sqrt{\frac{\sum_{i=1}^{5}(x_i - \overline{x})^2}{\sum_{i=1}^{5}(y_i - \overline{y})^2}}$$

$$= 0.255 \cdot \sqrt{\frac{0.40}{0.0259}} = 1.002$$

第7章要点

1、理解平均值的置信区间的概念及计算。

2、了解显著性检验的方法，会运用各种检验法进行显著性差异的检验。

3、掌握异常值的取舍方法并应用。

4、了解误差的传递，会计算极值误差和极值相对误差。

5、掌握回归分析法。

6、掌握消除系统误差和减小随机误差的方法。

第7章作业

P269 No.8, No14

P270 No.19, N0. 20, No.21

P271 No.28